International Journal of Engineering &Technology, *7 (1.8) (2018) 81-85*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Semi-supervised learning: a brief review

Y C A Padmanabha Reddy [1] *, P Viswanath [2], B Eswara Reddy [3]

[1] *Dept. of CSE, Research Scholar, JNTUA Ananthapuramu-515 002*
[2] *Dept. of CSE, IIIT Sricity, Chittoor-517588*
[3] *Dept of CSE, JNTUA college of Engineering, kalikiri-517234*
**Corresponding author E-mail: ananthayc.reddy@gmail.com*

Abstract

Most of the application domain suffers from not having sufficient labeled data whereas unlabeled data is available cheaply. To get labeled instances, it is very difficult because experienced domain experts are required to label the unlabeled data patterns. Semi-supervised learning addresses this problem and act as a half way between supervised and unsupervised learning. This paper addresses few techniques of Semi-supervised learning (SSL) such as self-training, co-training, multi-view learning, TSVMs methods. Traditionally SSL is classified in to Semi-supervised Classification and Semi-supervised Clustering which achieves better accuracy than traditional supervised and unsupervised learning techniques. The paper also addresses the issue of scalability and applications of Semi-supervised learning.

Keywords: *Semi-Supervised Learning; Labeled Data; Unlabeled Data; SSL Methods; Training Data; Test Data.*

1. Introduction

Semi-supervised learning (SSL) is more recent when com-pared with the supervised and unsupervised learning. Supervised learning is to learn from the set of given examples, where each example consists of the problem instance along with its label (usually given by some expert in that field). For example, in classification problem, the data element to be classified is represented as a feature vector and the class is given as a categorical label. The example set, which is also called the training set or the labeled set, is used to build the classifier which can be used to classify any new given data instance. In unsupervised learning, we are not provided with any labeled set. But, we are given with an unlabeled data. The primary task of unsupervised learning is to find the structure present in the data set, like the clustering structure. Unsupervised learning is more difficult when compared with supervised learning, since we do not have the ground truth to evaluate the results [1].

In most of the real-world application domains like image processing and text processing, where there is an abundant supply of unlabeled data, which requires human experts to label the unlabeled data, is an expensive task. Many real-world applications the data is very sparse labeled data. Semi-supervised learning lies in between supervised and unsupervised learning, where provided with a mixture of labeled and unlabeled data become a significant role in recent research.

Many semi-supervised classification methods and semi-supervised clustering methods are available in the Literature [2], [3]. Class of semi-supervised learning methods have been proposed for both generative and discriminative techniques. Expectation Maximization (EM) is one of the generative semi-supervised methods. Text classification models with an approach of EM are discussed by Nigam [4].

Generative semi-supervised technique depends on the distribution of the input data and can fail even when the input data is not matched with the classification task [5]. Discriminative semi-supervised methods [6] including the probabilistic and non-probabilistic methods, such as transductive support vector machines (TSVMs and S3VMs) and graph-based methods assume densities with the class. These methods can fail when the classes are strongly interleaved.

This paper addresses the scalability issue of SSL methods and also briefly tabulated the important real-world applications of SSL.

2. Machine learning approaches

Machine learning approaches are broadly categorized into two. i) Supervised Learning ii) Unsupervised Learning. Further these two categories are divided into two a) Semi-Supervised Learning b) Semi-Unsupervised Learning, is shown in fig 1.

3. Supervised learning

Supervised learning builds a knowledge base from the pre-classified patterns that supports to classify new patterns. The major task of this learning is to map the input features to an output called class. The outcome of this learning is to construct a model

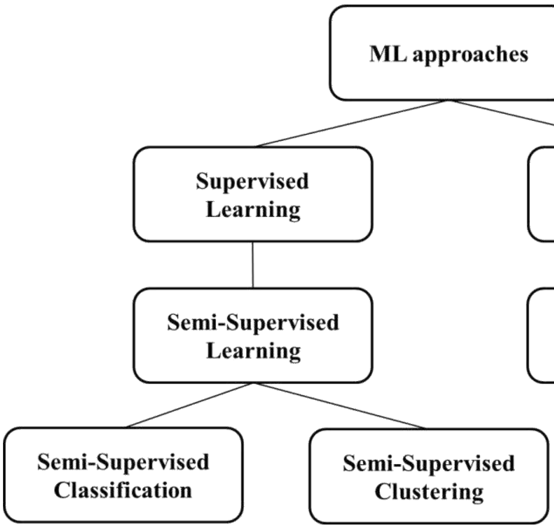
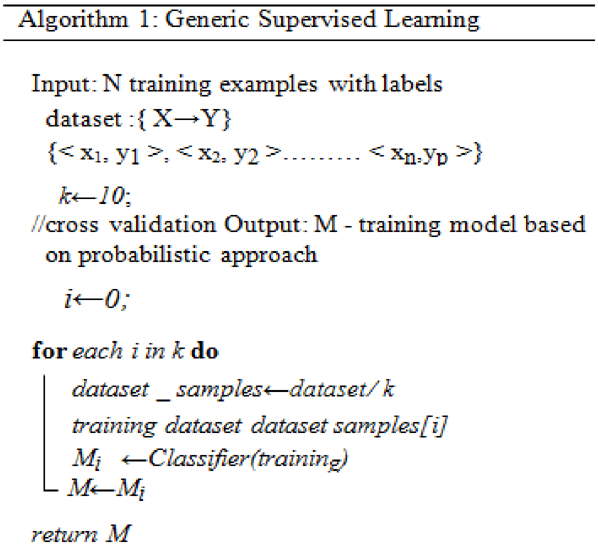


Fig. 1: Machine Learning Approaches.

By examining from input patterns. The model can be used to correctly classify unseen instances. In general, it can be represented as a function f(x) as input patterns and an output class y. The pre-classified patterns named as training set (TS) a pair wise input & output and an unseen pattern named as test set has only input patterns.

$$\text{Let}(DS) = \{< X_1, y_1 >, < X_2;\ y_2 > \cdots < Xn;\ yn >\}$$

Where n is the number of patterns or observation and p represents number of classes. The generic supervised learning Algorithm 1 is given below. There are several supervised learning algorithms are proposed like Decision Trees, bagging, Boosting, Random Forest, k-NN, Logistic Regression, Neural Networks, Support Vector Machines, Naive Base, Bayesian Networks.

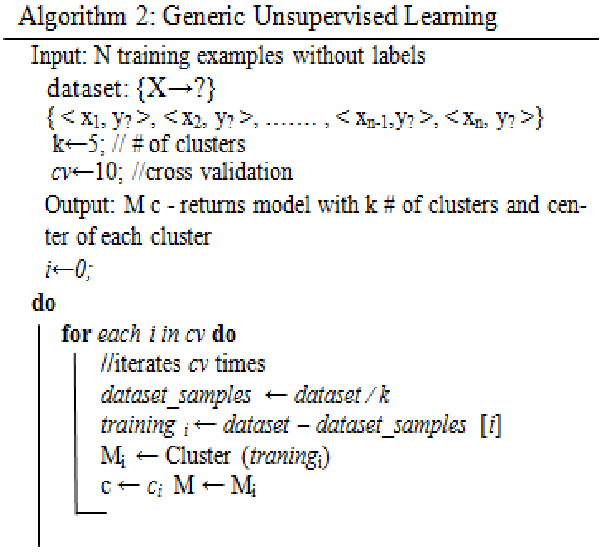


```
Algorithm 1: Generic Supervised Learning

Input: N training examples with labels
  dataset :{ X→Y}
  {< $x_1$, $y1$ >, < $x_2$, $y2$ >......... < $x_n$,$y_p$ >}
   $k \leftarrow 10$;
//cross validation Output: M - training model based
  on probabilistic approach
    $i \leftarrow 0$;

for each $i$ in $k$ do
    dataset _ samples←dataset/ k
    training dataset dataset samples[i]
    $M_i$  ←Classifier(training)
    $M \leftarrow M_i$

return M
```

4. Unsupervised clustering

Unsupervised learning studies how systems can learn to repre-sent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with supervised learning or reinforcement learning, there are no explicit target outputs or environmental evaluations associated with each input; rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output. The only things that unsupervised learning methods have to work with are the observed input patterns x_i, which are often assumed to be independent samples from an underlying unknown probability distribution $P_I[x]$, and some explicit or implicit a priori information as to what is important. The generic Unsupervised learning Algorithm 2 is given below.

Density estimation techniques explicitly build statistical models (Such as Bayesian Network) of how underlying causes could create the input. Feature extraction techniques try to extract statistical regularities (or sometimes irregularities) directly from the inputs. However unsupervised learning also encompasses many other techniques that seek to summarize and explain key features of the data. Many methods employed in unsupervised learning are based on data mining methods used to pre-process data. Examples of unsupervised learning algorithms are, clustering (k-means, mixture models, hierar- chical clustering), Expectation maximization algorithm (EM), Principal component analysis (PCA), Independent component analysis (ICA), Singular value decomposition (SVD)



```
Algorithm 2: Generic Unsupervised Learning
Input: N training examples without labels
  dataset: {X→?}
  { < $x_1$, $y_?$ >, < $x_2$, $y_?$ >, ....... , < $x_{n-1}$,$y_?$ >, < $x_n$, $y_?$ >}
   k←5; // # of clusters
   cv←10; //cross validation
  Output: M c - returns model with k # of clusters and cen-
  ter of each cluster
  $i \leftarrow 0$;
do
    for each i in cv do
        //iterates cv times
        dataset_samples ← dataset / k
        training $_i$ ← dataset – dataset_samples [i]
        $M_i$ ← Cluster (traning$_i$)
        c ← $c_i$ M ← $M_i$
```

While till all training examples assigns clusters return Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with supervised learning or reinforcement learn-ing, there are no explicit target outputs or environmental evaluations associated with each input; rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output. The only things that unsupervised learning methods have to work with are the observed input patterns x_i, which are often assumed to be independent samples from an underlying unknown probability distribution P_I [x], and some explicit or implicit a priori information as to what is important. The generic Unsupervised learning Algorithm 2 is given below.

While till all training examples assigns clusters return M, c.

5. Semi-supervised learning

Semi-supervised learning (SSL) is a type of Machine Learning (ML) technique. It is half-way between supervised and unsupervised learning i. e the dataset is partially labeled is shown in Figure 2. The main objective of SSL is to overcome the drawbacks of both supervised and unsupervised learning. Supervised learning requires huge amount of training data to classify the test data, which is cost effective and time consuming process. On the other hand, unsupervised learning doesn’t require any labeled data, which clusters the data based on similarity in the data points by using either clustering or maximum likelihood approach. The main downfall of this approach, it can’t cluster an unknown data accurately. To overcome these issues, SSL has been proposed by research community, which can learn with small amount of training data can label the unknown (or) test data. SSL builds a model with few labeled patterns as training data and treats the rest of the patterns as test data. The generic Semi-supervised learning Algorithm 3 is given below.

Semi-supervised learning is further dived into two types i) Semi-Supervised classification and ii) Semi-Supervised Clustering is discussed in the below section.

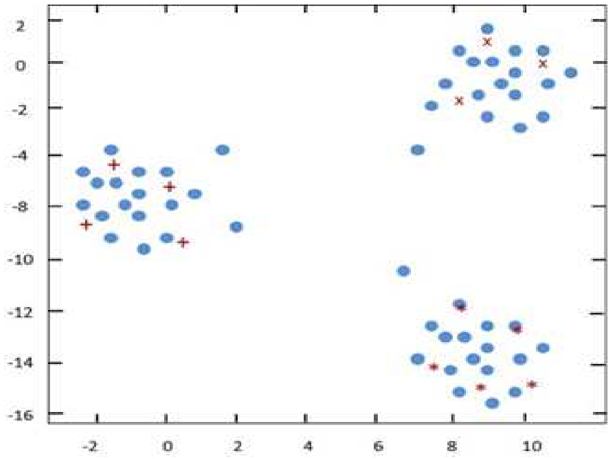
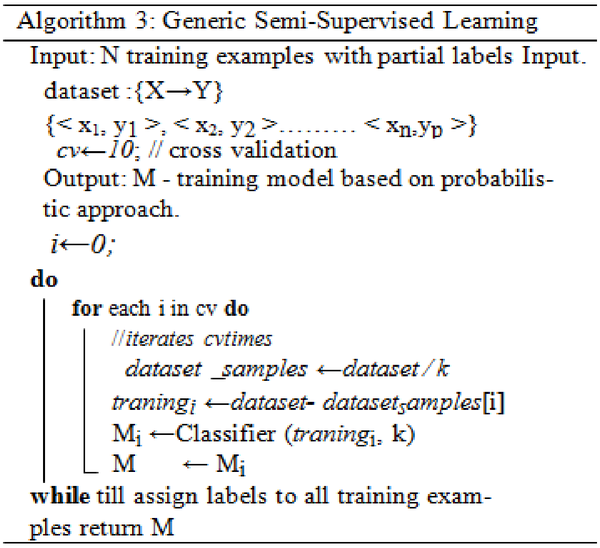


Fig. 2: Semi-Labeled Dataset.

5.1. Semi-Supervised Classification

Semi-Supervised Classification (SSC) is similar to Super-vised approach, require more training data to classify the test data. But in SSC, use less train data to classify the large amount of test data. By using this semi supervised classification we reduce the usage of the training data. Currently, more unlabeled data patterns are available sufficiently in the research community but the labeled data is not available. Because to design of the training data is cost effective and time consuming [2].



```
Algorithm 3: Generic Semi-Supervised Leaming
Input: N training examples with partial labels Input.
  dataset :{X→Y}
  {< x1, y1 >, < x2, y2 >......... < xn,yp >}
   cv←10; // cross validation
  Output: M - training model based on probabilis-
  tic approach.
   i←0;
do
    for each i in cv do
        //iterates cvtimes
          dataset _samples ←dataset / k
        traning_i ←dataset- dataset_samples[i]
        M_i ←Classifier (traning_i, k)
        M      ← M_i
while till assign labels to all training exam-
ples retum M
```

In [7] authors proposed an approximation solution to label the test patterns by selective incremental transductive nearest neighbor (NN) classifier (SI-TNNC). Authors has compared their results with 5 diversified datasets and 5 different algorithms and shown 3 out of 5 algorithms, the SI-TNNC has higher accuracy compared to standard algorithms like ID3 and 3NN etc.

In [8] proposed a framework to classify partially labeled data that improve the classification accuracy. Authors main idea was that classification defined only on the sub-main-fold rather than ambient space. The proposed algorithm uses adjacency graph for approximation of labels. The framework main uses Laplace-Beltrami operator which produces Hilbert space on sub-main-fold. To accomplish this task, the frame-work requires only unlabeled examples.

Real-time traffic classification using semi-supervised learning has proposed in [9]. Anthers has proved that semi-supervised learning is always better that the supervised learning in terms of preparation of training data and to build a model. They had successfully used SSL on network traffic classification of various networking applications in real-time.

5.2. Semi-supervised clustering

Semi supervised clustering is a special case of clustering. Generally in clustering we use unlabeled data patterns for clustering. But in semi supervised clustering we use both labeled and unlabeled data with side information as pair wise (must-link and cannot link) constraints which helps to cluster the data patterns [2].

Semi-supervised Single Link (SSL) cluster approach solves the problem of arbitrary shaped cluster in [20]. SSL overcomes noisy-bridge problem which is a distance between clusters by considering the predefined distance matrix with minimal constraints. Authors proven their results on both synthetic and real world datasets

Self-training approach is usually applied technique in SSL. In this approach, the algorithm classify with few labeled training data, then it classify an unlabeled data and these predicted patterns then added to the training set. The process is repeated until the test set empty [21]. Few algorithms try to bypass by “unlearn”, an unlabeled points if the predicted data patterns are below threshold. Self-training has been applied in applications like natural language processing (NLP) tasks.

Support vector Machines is a standard classifier where the labeled data only used, whereas unlabeled data is used in TSVMs [21]. TSVMs is the extension for SVM, with the objective is to provide labels to the unlabeled data patterns, which can linear separable that can maximize the margin on both the original labeled data patterns from unlabeled data. TSVMs became more popular and used in many applications such as image retrieval, bio-informatics and for named entity recognition.

A probabilistic framework has proposed in [22] for semi-supervised clustering. Authors minimized the objective func-tion derived from the posterior energy of the Hidden Markov Random Fields (HMRF). Their framework demonstrated on several text data sets shown that advantage of semi-supervised learning.

6. Scalability issues of semi-supervised learning methods & applications

SSL methods doesn't scale well for the large amount of data [1] especially, graph based semi-supervised methods takes cubic time complexity O (n^3). Speed-up improvements have been proposed (Mahdaviani et al. 2005; Delalleau et al. 2005; Zhu and Lafferty 2005; Yu et al. 2005; Garcke and Griebel 2005; and more), their effectiveness has yet to be proven on real large problems [15] In many supervised learning papers, SSL methods have not addressed large scale problems. The unlabeled dataset size is huge in terms of patterns, to handling the huge amount of unlabeled data is a challenging task.

6.1. Applications

Table I describes several semi-supervised learning applications. Majority of the applications focused on accuracy using SSL. Further it can be expand to other areas like Spatial mining, Natural language processing, large volume of datasets like network traffic, Speech recognition etc. Most of the applications focuses on accuracy as one of the metric to determine the performance of algorithm. But, accuracy may not suffice to classify patterns using semi-supervised learning. Further, we need to focus on other metrics like recall/precision etc. as well.

Table 1: Applications of Semi-Supervised Approach

Sl. No	Author Names	Methodology	Objective	Applications	Reference
1	YuZhou, Anlong Ming	Semi-Multiple Instance Learning.(Semi-MIL)	To achieve better accuracy	Objects are tracking	[10]
2	LeYao, Zhiqiang	SS-HELM (Semi-supervised Process data with Extreme Learning Machine	To model best soft sensors	Effective use of soft sensors in Industries.	[11]
3	Vivek Mighani and Richard Ribon Fletcher	Semi-supervised deep learning algorithm	To diagnoses the pulmonary disease.	Medical field: primary care and general patient monitoring	[12]
4	Bo-Hao chen et.al.	Semi-supervised algorithm	To remove noise images.	Image Categorization	[13], [14]
5	Ahmed et .al	Fuzzy Spectral Clustering	To get improved classification accuracy.	Hyper-spectral Image Classification	[15], [16]
6	Xinxing Xu; Wen Li	Multi-view weakly labeled learning.	To label the unlabeled text, documents.	Text Classification	[17]
7	Erman et. al.	Semi-supervised learning	Off-line/real-time traffic classification	Network traffic classification	[9]
8	Helmut Grabner et. al.	Semi-supervised	Identify tracking failure (drifting)	On-Line Boosting for Robust Tracking	[18]
9	Maria-Florina Balcan et. al.	Semi-supervised Learning in Web-cam images	To identify web-cam Images.	Person Identification in Web-cam Images	[19]

7. Conclusion

Semi-supervised learning addresses the issues of supervised and unsupervised approaches. An unlabeled data patterns along with labeled data pattern gives better accuracy when compared with supervised and unsupervised. This paper ad-dresses few SSL approach such as self-training, co-training, multi-view learning, TSVMs are addressed briefly. The understanding of standard existing methods and how these are related in SSL. In this paper, scalability is one of challenging issue has been addressed. SSL is the young discipline where the selection of good problem structure is very important to improve the performance. Further, we work around scalability and other metrics to classify patterns using SSL

Acknowledgement

This work is partly supported by a UGC-SERO Minor Research Project with Reference No. MRP-4609/14.

References

[1] A. Jain, M. Murty, and P. Flynn, "Data clusterting: A review acm computing surveys, vol. 31," 1999.

[2] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," IEEE Transactions on Neural Networks, vol. 20, no. 3, pp. 542–542, 2009. https://doi.org/10.1109/TNN.2009.2015974.

[3] P. K. Mallapragada, University, Some contributions to semi-supervised learning. Michigan State 2010.

[4] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using e," Machine learning, vol. 39, no. 2, pp. 103–134, 2000. https://doi.org/10.1023/A:1007692713085.

[5] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, "Generative or discriminative? Getting the best of both worlds," Bayesian Stat, vol. 8, no. 3, pp. 3–24, 2007.

[6] R. K. Ando and T. Zhang, "Two-view feature generation model for semi-supervised learning," in Proceedings of the 24th international conference on Machine learning. ACM, 2007, pp. 25–32. https://doi.org/10.1145/1273496.1273500.

[7] P. Viswanath, K. Rajesh, C. Lavanya, and Y. P. Reddy, "A selective incremental approach for transductive nearest neighbor classification," in Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE. IEEE, 2011, pp. 221–226.

[8] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," Machine learning, vol. 56, no. 1-3, pp. 209–239, 2004. https://doi.org/10.1023/B:MACH.0000033120.25363.1e.

[9] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Offline/realtime traffic classification using semi-supervised learning," Per-formance Evaluation, vol. 64, no. 9, pp. 1194–1213, 2007. https://doi.org/10.1016/j.peva.2007.06.014.

[10] C. Methani, R. Thota, and A. Kale, "Semi-supervised multiple instance learning based domain adaptation for object detection," in Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing. ACM, 2012, p. 13. https://doi.org/10.1145/2425333.2425346.

[11] L. Yao and Z. Ge, "Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application," IEEE Transactions on Industrial Electronics, vol. 65, no. 2, pp. 1490–1498, 2018. https://doi.org/10.1109/TIE.2017.2733448.

[12] D. Chamberlain, R. Kodgule, D. Ganelin, V. Miglani, and R. R. Fletcher, "Application of semi-supervised deep learning to lung sound analysis," in Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the. IEEE, 2016, pp. 804–807. https://doi.org/10.1109/EMBC.2016.7590823.

[13] L. Liu, L. Chen, C. P. Chen, Y. Y. Tang et al., "Weighted joint sparse representation for removing mixed noise in image," IEEE

transactions on cybernetics, vol. 47, no. 3, pp. 600–611, 2017. https://doi.org/10.1109/TCYB.2016.2521428.

[14] B.-H. Chen, J.-L. Yin, and Y. Li, “Image noise removing using semi-supervised learning on big image data,” in Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on. IEEE, 2017, pp. 338–345. https://doi.org/10.1109/BigMM.2017.42.

[15] P. S. S. Aydav and S. Minz, “Modified self-learning with clustering for the classification of remote sensing images,” Procedia Computer Science, vol. 58, pp. 97–104, 2015. https://doi.org/10.1016/j.procs.2015.08.034.

[16] J. Zhang, Y. Han, J. Tang, Q. Hu, and J. Jiang, “Semi-supervised image-to-video adaptation for video action recognition,” IEEE transactions on cybernetics, vol. 47, no. 4, pp. 960–973, 2017. https://doi.org/10.1109/TCYB.2016.2535122.

[17] X. Xu, W. Li, D. Xu, and I. W. Tsang, “Co-labeling for multi-view weakly labeled learning,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 6, pp. 1113–1125, 2016. https://doi.org/10.1109/TPAMI.2015.2476813.

[18] H. Grabner, C. Leistner, and H. Bischof, “Semi-supervised on-line boosting for robust tracking,” Computer Vision–ECCV 2008, pp. 234– 247, 2008. https://doi.org/10.1007/978-3-540-88682-2_19.

[19] M.-F. Balcan, A. Blum, P. P. Choi, J. D. Lafferty, B. Pantano, M. R. Rwebangira, and X. Zhu, “Person identification in webcam images: An application of semi-supervised learning,” 2005.

[20] Y. P. Reddy, P. Viswanath, and B. E. Reddy, “Semi-supervised single-link clustering method,” in Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on. IEEE, 2016, pp. 1–5. https://doi.org/10.1109/ICCIC.2016.7919689.

[21] T. Joachims, “Transductive inference for text classification using support vector machines,” in ICML, vol. 99, 1999, pp. 200–209.

[22] S. Basu, M. Bilenko, and R. J. Mooney, “A probabilistic framework for semi-supervised clustering,” in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’04. New York, NY, USA: ACM, 2004, pp. 59–68. [Online]. Available.